A short survey on protein blocks.

Agnel Joseph, Garima Agarwal, Swapnil Mahajan, Jean-Christophe Gelly, Lakshmipuram S. Swapna, Bernard Offmann, Frédéric Cadet, Aurélie Bornot, Manoj Tyagi, Hélène Valadié, et al.

A short survey on protein blocks

Agnel Praveen Joseph [1 +], Garima Agarwal [2 +], Swapnil Mahajan [2 3], Jean-Christophe Gelly [1], Lakshmipuram S. Swapna [2], Bernard Offmann [5 4], Frédéric Cadet [4 5], Aurélie Bornot [1], Manoj Tyagi [6], Hélène Valadié [7], Bohdan Schneider [8], Catherine Etchebest [1], Narayanaswamy Srinivasan [2], Alexandre G. De Brevern [1 4 *]

[1] *DSIMB, Dynamique des Structures et Interactions des Macromolécules Biologiques Université Paris-Diderot - Paris VII , INTS , INSERM : U665 , INTS, 6 rue Alexandre Cabanel, 75739 Paris Cedex 15 FRANCE,FR*

[2] *Molecular Biophysics Unit Indian Institute of Science , Bangalore 560 012,IN*

[3] *NCBS, National Centre for Biological Sciences Tata institute of Fundamental Research , UAS, GKVK Campus, Bellary Road, Bangalore 560 065,IN*

[4] *Protéines de la membrane érythrocytaire et homologues non-érythroides INSERM : U665 , Université Paris-Diderot - Paris VII , INTS , INTS 6, Rue Alexandre Cabanel 75739 PARIS CEDEX 15,FR*

[5] *Faculté des Sciences et Technologies Université de la Réunion , 15 Avenue René Cassin, BP 7151, 97715 Saint Denis Messag Cedex 09, La Réunion,FR*

[6] *Computational Biology Branch NLM , NCBI , 8600 Rockville Pike, Bethesda, MD 20894,US*

[7] *LPCV, Laboratoire de physiologie cellulaire végétale CNRS : UMR5168 , INRA : UR1200 , CEA : DSV/IRTSV , Université Joseph Fourier - Grenoble I , bat. C2 17 Rue des martyrs 38054 GRENOBLE CEDEX 9,FR*

[8] *Institute of Biotechnology AS CR Institute of Biotechnology AS CR , Prague,CZ*

* Correspondence should be adressed to: Alexandre De Brevern <alexandre.debrevern@univ-paris-diderot.fr >

+ The first two authors contributed equally to this article.

Abstract

Protein structures are classically described in terms of secondary structures. Even if the regular secondary structures have relevant physical meaning, their recognition from atomic coordinates has some important limitations such as uncertainties in the assignment of boundaries of helical and β-strand regions. Further, on an average about 50% of all residues are assigned to an irregular state, i.e ., the coil. Thus different research teams have focused on abstracting conformation of protein backbone in the localized short stretches. Using different geometric measures, local stretches in protein structures are clustered in a chosen number of states. A prototype representative of the local structures in each cluster is generally defined. These libraries of local structures prototypes are named as " structural alphabets". We have developed a structural alphabet, named Protein Blocks, not only to approximate the protein structure, but also to predict them from sequence. Since its development, we and other teams have explored numerous new research fields using this structural alphabet. We review here some of the most interesting applications.

Author Keywords protein structures ; biochemistry ; amino acids ; secondary structures ; propensities ; structural alphabet ; structure prediction ; structural superimposition ; mutation ; binding site ; Bayes theorem ; Support Vector Machines.

Introduction

Protein structures have been classically described in two regular states (α-helix and β-strand) and the remaining unassigned regions as an irregular state (coil), this state correspond to a large number of diverse conformations. Nonetheless, the use of only three states oversimplifies the description of protein structures. A detailed description for 50% of the residues classified as coils is missed even when they encompass repeating local structure. Description of local protein structures have hence focused on the elaboration of complete sets of small prototypes or "structural alphabets" (SAs), that help to approximate every part of the protein backbone (Offmann, et al., 2007). Designing a structural alphabet requires identification of a set of average recurrent local protein structures that (efficiently) approximates every part of known structures. As each residue is associated to one of these prototypes, the whole 3D protein structure can be translated into a series of prototypes (letters) in 1D, as the sequence of prototypes.

Figure 1 gives an example of encoding of a protein structures with a Structural Alphabet. The N-terminal extremity of Aspergillus niger acid phosphatase (Kostrewa, et al., 1999) chain B is shown. To each residue, a local protein structure prototype was associated. Thus, the coil region could be precisely described as a succession of small protein prototypes instead of a succession of identical states.

Protein Blocks

Secondary structure assignments are widely used to analyze protein structures. However, it often gives a coarse description of 3D protein structures, with about half of the residues being assigned to an undefined state (Bornot and de Brevern, 2006). Moreover, the structural diversity observed in α-helices and β-strands, is hidden. Indeed, α-helices are frequently not linear, and are either curved (58%)

or kinked (17%) (Martin, et al., 2005). The absence of secondary structure assignment for a significant proportion of the residues has led to the development of local protein structure libraries that are able to approximate all (or almost all) of the local protein structures without using classical secondary structures. These libraries yielded prototypes that are representative of local folds found in proteins. The complete set of local structure prototypes defines a structural alphabet (Offmann, et al., 2007).

Ten years ago, Pr. Serge Hazout developed a novel structural alphabet with two specific goals (de Brevern, et al., 2000): (i) to obtain a good local structure approximation and (ii) to predict local structures from sequence. Fragments that are five residues in length were coded in terms of the φ/ϕ dihedral angles. A Root Mean Square Deviation on Angle (RMSDA) score was used to quantify the structural difference among the fragments (Schuchhardt, et al., 1996). Using an unsupervised cluster analyser related to Self Organizing Maps (SOM (Kohonen, 1982 ; Kohonen, 2001)), a three step training process was carried out. The first step involved learning of structural difference of fragments in terms of RMSDA and in the second step the transition probability (probability of transition from one fragment to another in a sequence) was also considered along with the RMSDA, i.e ., in a similar way to Markov model (Rabiner, 1989). In the third step, the constraint based on transition probability was removed. Optimal prototypes were identified by considering both the structural approximation and the prediction rate. A set of 16 prototypes called Protein Blocks (PBs), represented as average dihedral vectors, was obtained at the end of this process (de Brevern, et al., 2000).

These PBs are displayed represented in Figure 2 . The PBs m and d can be described roughly as prototypes for central α-helix and central β-strand, respectively. PBs a through c primarily represent β-strand N-caps and PBs e and f , β-strand C-caps; PBs g through j are specific to coils, PBs k and l to α-helix N-caps, and PBs n through p to α-helix C-caps. This structural alphabet allows a good approximation of local protein 3D structures with a root mean square deviation (rmsd) now evaluated at 0.42 Å on average (de Brevern, 2005). PBs have been assigned using in-house software (available at http://www.dsimb.inserm.fr/DOWN/LECT/) or using PBE web server (http://bioinformatics.univ-reunion.fr/PBE/) (Tyagi, et al., 2006).

PBs (de Brevern, et al., 2000) have been used both to describe the 3D protein backbones (de Brevern, 2005) and to perform local structure prediction(de Brevern, et al., 2007 ; de Brevern, et al., 2000 ; de Brevern, et al., 2002 ; Etchebest, et al., 2005). Our earlier work on PBs have shown that PBs are effective in describing and predicting conformations of long fragments (Benros, et al., 2006 ; Benros, et al., 2009 ; Bornot, et al., 2009 ; de Brevern, et al., 2007 ; de Brevern and Hazout, 2001 ; de Brevern and Hazout, 2003 ; de Brevern, et al., 2002) and short loops (Fourrier, et al., 2004 ; Tyagi, et al., 2009 ; Tyagi, et al., 2009), analyzing protein contacts (Faure, et al., 2008), in building a transmembrane protein (de Brevern, 2005 ; de Brevern, et al., 2009), and in defining a reduced amino acid alphabet to aid design of mutations (Etchebest, et al., 2007). This reduced amino acid alphabet was recently proved suitable for predicting protein families or sub-families and secretory proteins of P. falciparum (Zuo and Li, 2009 ; Zuo and Li, 2009). We have also used protein blocks to superimpose and to compare protein structures (Tyagi, et al., 2008 ; Tyagi, et al., 2006 ; Tyagi, et al., 2006).

Other laboratories have taken advantage of PBs to reconstruct globular protein structures (Dong, et al., 2007), design peptides (Thomas, et al., 2006) and to define binding site signatures (Dudev and Lim, 2007). Novel prediction methodologies (Li, et al., 2009 ; Rangwala, et al., 2009 ; Zimmermann and Hansmann, 2008) and fragment-based local statistical potentials (Li, et al., 2009) were also developed. The features of this alphabet have been compared by Karchin et al. (Karchin, et al., 2003) with those of 8 other structural alphabets showing that our PB alphabet is highly informative, with the best predictive ability of those tested. Among the available SAs, it is the most widely used SA today.

Applications

Binding site signature

PBs enable the detection of structural similarity between proteins with excellent efficiency. Dudev, Lim and co-workers, (Dudev and Lim, 2001 ; Yang, et al., 2008), used this concept to locate structural motifs of metal/ligand-binding sites in proteins (Dudev and Lim, 2007). They encoded a protein structure databank in terms of PBs and they located PBs encompassing specific metal-binding sites. Then, a discontinuous PB pattern, similar to a PROSITE pattern, was defined. First, the structural motifs of Cys_4 Zn-finger domains, which are known to adopt a specific structure, have been analyzed. Then, they focused on structural motifs of Mg^{2+} -binding sites in a set of non-redundant Mg^{2+} -binding proteins. Four Mg^{2+} -structural motifs that showed important relationships between them were identified. Other features of the proteins were also defined (Dudev and Lim, 2007). This strategy can be easily extended to other cases. Recently, they have extended the approach to DNA and RNA binding sites, highlighting a novel non-specific motif enabling diverse interactions with DNA and RNA as with proteins (Wu, et al., 2010).

Definition of a reduced amino acid alphabet

Reduced amino acid alphabet is a popular concept that is explored by many research teams. Indeed, the appropriate selection of an amino acid type in a reliable set is particularly helpful to limit the number of experiments. Most of approaches were mainly based on sequence properties, i.e ., (Akanuma, et al., 2002 ; Clarke, 1995 ; Kamtekar, et al., 1993).

In this area, PBs not only help in describing protein structures, but are also useful in extraction of sequence – structure relationships. Based on this relation, we proposed association of amino acids in a limited number of clusters. This approach permits an exchange of amino acids which are equivalent in terms of sequence – structure relationship, while maintaining local protein structure conformation (Etchebest, et al., 2007). Zuo and Li used this reduced amino acid alphabet to predict different properties through a learning approach (Zuo and Li, 2009 ; Zuo and Li, 2009).

Long structural fragments

PBs are 5 residues long fragments. To assess the structural stability of these short fragments, we identified the most frequent series of 5 consecutive PBs which are 9 residues long. Then, we selected 72 most frequent series and named them Structural Words (SWs). Interestingly, SWs encompass 92% of the residues (nearly 100% of the repetitive structures and 80% of secondary structure coil). By using most of the SWs, it was possible to create a simple network describing most of the transitions between the SWs in proteins (de Brevern, et al., 2002). SWs yield a pertinent description of a large part of 3D structures, but as they constitute a sub-set of all five PBs combinations, they do not allow a description for every part of the protein structures. So, we have developed a novel approach named Hybrid Protein Model (HPM (de Brevern and Hazout, 2000)). This innovative approach made it possible to create longer prototypes comprising 10 to 13 residues. (Benros, et al., 2006 ; Benros, et al., 2003 ; Benros, et al., 2009 ; Benros, et al., 2002 ; de Brevern and Hazout, 2001 ; de Brevern and Hazout, 2003). This resulted in higher structural variability for the longer fragments through a significant increase in the number of prototypes, e.g ., 100 to 130 prototypes (Benros, et al., 2006 ; Benros, et al., 2009 ; de Brevern and Hazout, 2001 ; de Brevern and Hazout, 2003). These longer fragments were used to perform structural superimposition (de Brevern and Hazout, 2001), methodological optimisation (Benros, et al., 2003 ; de Brevern and Hazout, 2003), and analysis of sequence – structure relationship (Benros, et al., 2006 ; Benros, et al., 2009 ; Bornot, et al., 2009 ; de Brevern and Hazout, 2001). A modified version of HPM proposed by Pr. Serge Hazout, has led to the construction of networks of local protein structures (Hazout, 2005).

Structural alignment

The structural alphabet allows translating protein three-dimensional structures into a series of letters (see Figure 1a). Consequently, it is possible to use classical sequence alignment methodology to perform structure-based alignment (see Figure 1b). The main difficulty lies in obtaining a pertinent substitution matrix, to find the similarity score between PBs for alignments. Using the homologues of known 3-D structure in PALI database (Gowri, et al., 2003) encoded in terms of PBs, a PB substitution matrix was computed (Tyagi, et al., 2006). A dedicated webserver has been developed (http://bioinformatics.univ-reunion.fr/PBE/) that performs optimal alignments of a query protein structure with entries of 3-D structures in a database, using PBs and the substitution matrix [18]. A recent benchmark has proved that this method is most efficient in mining PDB and identifying proteins with similar 3-D structure (Tyagi, et al., 2008).

From this work, new developments have been performed. A first one directly relates to the use of substitution matrix and concerns the characterization of conformational patterns in active and inactive forms of kinases. Comparison of closely related kinases indicates a higher global similarity between the active state kinases compared to inactive states (as reflected from their PB scores) (Agarwal, et al., 2010). The second axis focuses on the database, which is the basis for generating the substitution matrix, i.e ., PALI database. The superimposed structures of PALI show regions with correct alignments linked with regions more difficult to align (named variable regions). A novel optimisation of the superposition based on PBs, shows a global improvement in the variable regions. Hence, PBs improve PALI database alignments (Agarwal et al ., in preparation). The last axis concerns the alignment approach. Even though the recent benchmark has proved the quality of the methodology (Tyagi, et al., 2008), some structural alignmentsshow poor consistency. Optimisation of the substitution matrix and a novel alignment methodology improved both the mining and the superimposition of protein structures (Joseph et al ., in preparation). Figure 3 illustrates a very difficult case of superimposition of the threedimensional structures of Aspergillus niger acid phosphatase (Kostrewa, et al., 1999) and Escherichia coli periplasmic glucose-1-phosphatase (Lee, et al., 2003). The superimposition obtained with our previous approach (Tyagi, et al., 2006) is quite poor. Our novel procedure allows the recognition of highly similar regions (shown on Figure 3e) and gives a spectacular improvement in the superimposition. Figures 3c and 3d show that the bottom part of the protein structures can be truly superimposed. Figure 4 gives the alignment in terms of PBs.

Prediction

Like secondary structure prediction, it is possible to predict local structures in terms of structural alphabet (see Table 1 for a summary of all prediction approaches). Indeed, concomitant to an accurate local 3D structures description, definition of PBs was driven by prediction capabilities (de Brevern, et al., 2000). The prediction principle is based on Bayes' theorem. First, a set of protein chains used in training were encoded in terms of PBs, using minimal RMSDA criterion. Then, sequence windows of 15 residues length were considered for calculating the propensities associated with each PB. For each PB, the probability of occurrence of an amino acid at each position in the sequence window was calculated and an occurrence matrix was generated, i.e ., 16 for the 16 PBs. Bayes theorem was applied to predict the structure of new sequences. A prediction rate of 34.4% was achieved (de Brevern, et al., 2004 ; de Brevern, et al., 2000). Nonetheless, only one amino acid occurrence matrix is associated to each PB. Consequently, the sequence information is averaged. A clustering approach related to SOM (Kohonen, 1982 ; Kohonen, 2001) performed on PBs sequences revealed well-defined sequence

families for some PBs. For each sequence family, an amino acid occurrence matrix was then computed. This strategy increased sequence specificities for some PBs and permitted to achieve an improved prediction rate of 40.7% (de Brevern, et al., 2004 ; de Brevern, et al., 2000). Finally, a simulated annealing approach in the process of sequence family generation, helped to improve the overall prediction to 48.7% (Etchebest, et al., 2005). Importantly, this approach did not bring any biased or unbalanced improvements between the PBs. Combining the secondary structure information with the Bayesian prediction did not result in significant improvement of the prediction rate. A website, LocPred (http://www.dsimb.inserm.fr/~debrevern/LOCPRED/), which includes most of the tools developed so far, is available to perform these predictions (de Brevern, et al., 2004). Predictions were also performed with the SWs (de Brevern, et al., 2007 ; de Brevern, et al., 2002) and specifically for short loops (Fourrier, et al., 2004 ; Tyagi, et al., 2009).

A knowledge-based approach for predicting local backbone structure was also developed. In this case, overlapping fragments of 5 residues from a query sequence are extracted and queried against a pentapeptide database. In this database, which was built from SCOP database culled at 95% identity, each pentapeptide is mapped to a Protein Block. In absence of any “hit” in the database, pentapeptides in which constraint of identity in the central position (position 3) is relaxed are considered. Overall performance of the approach was about 62%.

Recent developments have been made by other teams. Li and co-workers who proposed an innovative approach for PB prediction, taking into account information from secondary structure and solvent accessibilities (Li, et al., 2009). Prediction rates were significantly improved (http://sg.ustc.edu.cn/lssrap/). Interestingly, this approach was found useful for fragment threading, pseudo sequence design, and local structure predictions.

Support Vector Machines methodology coupled with evolutionary information greatly improved the prediction rates. Hence, Zimmermann and Hansmann developed a method for PB prediction using SVMs with a radial basis function kernel, leading to an improvement of the prediction rate to 60–61% (Zimmermann and Hansmann, 2008). This method called Locustra is available online at http://www.fz-juelich.de/nic/cbb/service/service.php . In a very recent work, Rangwala, Kauffman and Karypis have developed a novel tool named svmPRAT (Rangwala, et al., 2009). It involves formulating the annotation problem as a classification or regression problem using support vector machines (http://www.cs.gmu.edu/~mlbio/prosat/). The use of such approach allows an impressive increase of prediction rate of about 69%. PB prediction is part of MONSTER (Minnesota prOteiN Sequence annoTation servER, http://bio.dtc.umn.edu/monster/). Thus, in less than a decade the prediction rate of PBs has doubled in a very efficient way.

As emphasized by Li and co-workers (Li, et al., 2009), it is often difficult to compare accurately the different studies because of different definitions considered for local structures, or different dataset and/or different criteria used for evaluating success predictions.

In order to extend the analyses to long structural fragments, HPM strategy was used to construct a new library of local structures. 120 structural clusters (named Local Structure Prototypes, LSPs) were then proposed to describe fragments that are 11-residue long (Benros, et al., 2006). An original prediction method based on logistic regressions was first developed for predicting local structures from a single sequence. This method proposed a short list of the best structural candidates among the 120 LSPs of the library. Considering a geometrical assessment, a prediction rate of 51.2 % was reached. This result was quite significant, given the fragment length and the high number of classes (Benros, et al., 2006). Recently, an improved prediction method based on SVMs and evolutionary information was proposed. A global prediction rate of 63.1% was achieved and prediction for 85% of proteins was improved. This method was shown to be among the most efficient of cuttingedge local structure prediction strategies (Bornot, et al., 2009).

Conclusions and Perspectives

Since 1989, nearly twelve different structural alphabets have been developed, e.g ., (Fetrow, et al., 1997 ; Ku and Hu, 2008 ; Sander, et al., 2006 ; Unger, et al., 1989 ; Unger and Sussman, 1993), for dedicated reviews see (Joseph, et al., 2010 ; Offmann, et al., 2007). However, almost none has been used outside their developing laboratories. PBs alphabet is the only exception. It is mainly due to the ease with which protein 3D structures can be encoded as PBs. It can be considered as the classical standard of Structural Alphabet as DSSP is the classical standard for secondary structure assignment (Kabsch and Sander, 1983).

PBs have been utilized in numerous different applications, e.g ., modelling of a transmembrane protein implicated in malarial infection (de Brevern, 2005 ; de Brevern, 2009 ; de Brevern, et al., 2009). It has also led to the development of an excellent superimposition method (Tyagi, et al., 2008) and is now used by various research teams (Joseph, et al., 2010). We have also developed confidence indexes associated to prediction accuracy (Bornot, et al., 2009 ; de Brevern, et al., 2000 ; Etchebest, et al., 2005). We now link the uncertainties with the prediction of protein flexibility, looking at data from X-ray analysis, Molecular Dynamics (Bornot et al ., submitted) and NMR. In the same way, superimposition approaches are currently improved. We also examine protein – protein interactions in the light of PBs (Swapna et al ., in preparation).

Ackowledgements:

The authors would like to thank the reviewers for their comments that help improve the manuscript. These works were supported by grants from the French Ministry of Research, University of Paris Diderot – Paris 7, University of Saint-Denis de la Réunion, French National Institute for Blood Transfusion (INTS), French Institute for Health and Medical Research (INSERM) and Indian Department of Biotechnology. APJ and GA are supported by CEFIPRA number 3903-E and Council of Scientific and Industrial Research, respectively. AB had a grant from the French Ministry of Research, MT has a post-doctoral fellowship from NIH and HV had a post-doctoral fellowship from CEA. NS and AdB acknowledge to CEFIPRA for collaborative grant (number 3903-E). BS and AdB acknowledge to Partenariat Hubert Curien Barrande (2010–2011).

References:

- Agarwal G , Dinesh D , Srinivasan N , de Brevern AG . Editor: Maulik U , Bandyopadhyay S , Wang J . 2010 ; Characterization of conformational patterns in active and inactive forms of kinases using Protein Blocks approach . Computational Intelligence and Pattern Analysis in Biological Informatics . Wiley ; in press
- Akanuma S , Kigawa T , Yokoyama S . 2002 ; Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set . Proc Natl Acad Sci U S A . 99 : 13549 - 13553
- Benros C , de Brevern AG , Etchebest C , Hazout S . 2006 ; Assessing a novel approach for predicting local 3D protein structures from sequence . Proteins . 62 : 865 - 880
- Benros C , de Brevern AG , Hazout S . 2003 ; Hybrid Protein Model (HPM): A Method For Building A Library Of Overlapping Local Structural Prototypes. Sensitivity Study And Improvements Of The Training . IEEE Workshop on Neural Networks for Signal Processing . 53 - 72
- Benros C , de Brevern AG , Hazout S . 2009 ; Analyzing the sequence-structure relationship of a library of local structural prototypes . J Theor Biol . 256 : 215 - 226
- Benros C , Hazout S , de Brevern AG . 2002 ; Extension of a local backbone description using a structural alphabet. “Hybrid Protein Model”: a new clustering approach for 3D local structures . International Workshop on Bioinformatics ISMIS . Lyon ; France 36 - 45
- Bornot A , de Brevern AG . 2006 ; Protein beta-turn assignments . Bioinformation . 1 : 153 - 155
- Bornot A , Etchebest C , de Brevern AG . 2009 ; A new prediction strategy for long local protein structures using an original description . Proteins . 76 : 570 - 587
- Clarke ND . 1995 ; Sequence ‘minimization’: exploring the sequence landscape with simplified sequences . Curr Opin Biotechnol . 6 : 467 - 472
- de Brevern AG . 2005 ; New assessment of a structural alphabet . In Silico Biol . 5 : 283 - 289
- de Brevern AG . 2009 ; New opportunities to fight against infectious diseases and to identify pertinent drug targets with novel methodologies . Infect Disord Drug Targets . 9 : 246 - 247
- de Brevern AG , Autin L , Colin Y , Bertrand O , Etchebest C . 2009 ; In silico studies on DARC . Infect Disord Drug Targets . 9 : 289 - 303
- de Brevern AG , Benros C , Gautier R , Valadie H , Hazout S , Etchebest C . 2004 ; Local backbone structure prediction of proteins . In Silico Biol . 4 : 381 - 386
- de Brevern AG , Etchebest C , Benros C , Hazout S . 2007 ; “Pinning strategy”: a novel approach for predicting the backbone structure in terms of protein blocks from sequence . J Biosci . 32 : 51 - 70
- de Brevern AG , Etchebest C , Hazout S . 2000 ; Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks . Proteins . 41 : 271 - 287
- de Brevern AG , Hazout S . 2000 ; Hybrid Protein Model (HPM): a method to compact protein 3D-structures information and physicochemical properties . IEEE - Computer Society . S1 : 49 - 54
- de Brevern AG , Hazout S . 2001 ; Compacting local protein folds with a “hybrid protein model” . Theo Chem Acc . 106 : 36 - 47
- de Brevern AG , Hazout S . 2003 ; ‘Hybrid protein model’ for optimally defining 3D protein structure fragments . Bioinformatics . 19 : 345 - 353
- de Brevern AG , Valadie H , Hazout S , Etchebest C . 2002 ; Extension of a local backbone description using a structural alphabet: a new approach to the sequencestructure relationship . Protein Sci . 11 : 2871 - 2886
- DeLano WLT . 2002 ; The PyMOL Molecular Graphics System . DeLano Scientific ; San Carlos, CA, USA http://www.pymol.org
- Dong QW , Wang XL , Lin L . 2007 ; Methods for optimizing the structure alphabet sequences of proteins . Comput Biol Med . 37 : 1610 - 1616
- Dudev M , Lim C . 2007 ; Discovering structural motifs using a structural alphabet: application to magnesium-binding sites . BMC Bioinformatics . 8 : 106 -
- Dudev T , Lim C . 2001 ; Modeling Zn2+-Cysteinate Complexes in Proteins . J Phys Chem . 105 : 10709 - 10714
- Etchebest C , Benros C , Bornot A , Camproux AC , de Brevern AG . 2007 ; A reduced amino acid alphabet for understanding and designing protein adaptation to mutation . Eur Biophys J . 36 : 1059 - 1069
- Etchebest C , Benros C , Hazout S , de Brevern AG . 2005 ; A structural alphabet for local protein structures: improved prediction methods . Proteins . 59 : 810 - 827
- Faure G , Bornot A , de Brevern AG . 2008 ; Protein contacts, inter-residue interactions and side-chain modelling . Biochimie . 90 : 626 - 639
- Fetrow JS , Palumbo MJ , Berg G . 1997 ; Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme . Proteins . 27 : 249 - 271
- Fourrier L , Benros C , de Brevern AG . 2004 ; Use of a structural alphabet for analysis of short loops connecting repetitive structures . BMC Bioinformatics . 5 : 58 -
- Gowri VS , Pandit SB , Karthik PS , Srinivasan N , Balaji S . 2003 ; Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database . Nucleic Acids Res . 31 : 486 - 488
- Hazout S . 2005 ; Une nouvelle méthode d'apprentissage: “Self-Learning by Information Share-Out” (SLISO) . Sixièmes Journées Ouvertes de Biologie, Informatique et Mathématiques (JOBIM) pour la génomique . Lyon 483 - 488
- Joseph AP , Bornot A , de Brevern AG . Editor: Rangwala H , Karypis G . 2010 ; Local Structure Alphabets . Protein Structure Prediction . wiley ; in press
- Kabsch W , Sander C . 1983 ; Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features . Biopolymers . 22 : 2577 - 2637
- Kamtekar S , Schiffer JM , Xiong H , Babik JM , Hecht MH . 1993 ; Protein design by binary patterning of polar and nonpolar amino acids . Science . 262 : 1680 - 1685
- Karchin R , Cline M , Mandel-Gutfreund Y , Karplus K . 2003 ; Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry . Proteins . 51 : 504 - 514
- Kohonen T . 1982 ; Self-organized formation of topologically correct feature maps . Biol Cybern . 43 : 59 - 69
- Kohonen T . 2001 ; Self-Organizing Maps . 3 Springer ;
- Kostrewa D , Wyss M , D'Arcy A , van Loon AP . 1999 ; Crystal structure of Aspergillus niger pH 2.5 acid phosphatase at 2.4 A resolution . J Mol Biol . 288 : 965 - 974
- Ku SY , Hu YJ . 2008 ; Protein structure search and local structure characterization . BMC Bioinformatics . 9 : 349 -
- Lee DC , Cottrill MA , Forsberg CW , Jia Z . 2003 ; Functional insights revealed by the crystal structures of Escherichia coli glucose-1-phosphatase . J Biol Chem . 278 : 31412 - 31418
- Li Q , Zhou C , Liu H . 2009 ; Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities . Proteins . 74 : 820 - 836
- Martin J , Letellier G , Marin A , Taly JF , de Brevern AG , Gibrat JF . 2005 ; Protein secondary structure assignment revisited: a detailed analysis of different assignment methods . BMC Struct Biol . 5 : 17 -
- Offmann B , Tyagi M , de Brevern AG . 2007 ; Local Protein Structures . Current Bioinformatics . 3 : 165 - 202
- Rabiner LR . 1989 ; A tutorial on hidden Markov models and selected application in speech recognition . Proceedings of the IEEE . 77 : 257 - 286
- Rangwala H , Kauffman C , Karypis G . 2009 ; svmPRAT: SVM-based protein residue annotation toolkit . BMC Bioinformatics . 10 : 439 -
- Sander O , Sommer I , Lengauer T . 2006 ; Local protein structure prediction using discriminative models . BMC Bioinformatics . 7 : 14 -
- Schuchhardt J , Schneider G , Reichelt J , Schomburg D , Wrede P . 1996 ; Local structural motifs of protein backbones are classified by self-organizing neural networks . Protein Eng . 9 : 833 - 842
- Thomas A , Deshayes S , Decaffmeyer M , Van Eyck MH , Charloteaux B , Brasseur R . 2006 ; Prediction of peptide structure: how far are we? . Proteins . 65 : 889 - 897

- Tyagi M , Bornot A , Offmann B , de Brevern AG . 2009 ; Analysis of loop boundaries using different local structure assignment methods . Protein Sci . 18 : 1869 - 1881
- Tyagi M , Bornot A , Offmann B , de Brevern AG . 2009 ; Protein short loop prediction in terms of a structural alphabet . Comput Biol Chem . 33 : 329 - 333
- Tyagi M , de Brevern AG , Srinivasan N , Offmann B . 2008 ; Protein structure mining using a structural alphabet . Proteins . 71 : 920 - 937
- Tyagi M , Gowri VS , Srinivasan N , de Brevern AG , Offmann B . 2006 ; A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications . Proteins . 65 : 32 - 39
- Tyagi M , Sharma P , Swamy CS , Cadet F , Srinivasan N , de Brevern AG , Offmann B . 2006 ; Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet . Nucleic Acids Res . 34 : W119 - 123
- Unger R , Harel D , Wherland S , Sussman JL . 1989 ; A 3D building blocks approach to analyzing and predicting structure of proteins . Proteins . 5 : 355 - 373
- Unger R , Sussman JL . 1993 ; The importance of short structural motifs in protein structure analysis . J Comput Aided Mol Des . 7 : 457 - 472
- Wu CY , Chen YC , Lim C . 2010 ; A structural-alphabet-based strategy for finding structural motifs across protein families . Nucleic Acids Res . in press
- Yang TY , Dudev T , Lim C . 2008 ; Mononuclear versus binuclear metal-binding sites: metal-binding affinity and selectivity from PDB survey and DFT/CDM calculations . J Am Chem Soc . 130 : 3844 - 3852
- Zimmermann O , Hansmann UH . 2008 ; LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach . J Chem Inf Model . 48 : 1903 - 1908
- Zuo YC , Li QZ . 2009 ; Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids . Amino Acids .
- Zuo YC , Li QZ . 2009 ; Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet . Peptides . 30 : 1788 - 1793

Figure 1

Principe of encoding of protein structures using Structural Alphabet

The N terminal extremity of chain B of Aspergillus niger acid phosphatase (Kostrewa, et al., 1999) (a) is encoded in terms of a structural alphabet (b). Each residue is approximated by a specific prototype, here a Protein Block. Hence, the crude description as a coil region (done by any secondary structure assignment method) is replaced by a more precise series of PBs dcbdfkl .

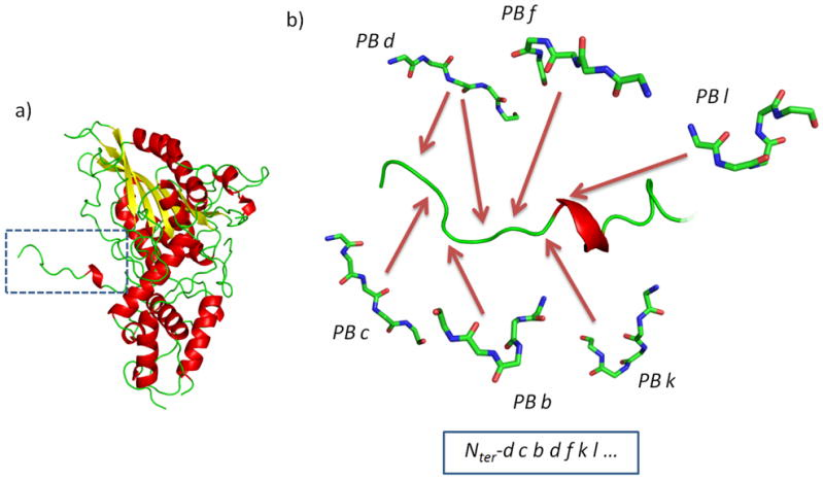

Figure 2
The Protein Blocks
PBs from a to p are shown using PyMol software (DeLano, 2002).. For each PB, the N cap extremity is on the left and the C-cap on the right. Each prototype is five residues in length and corresponds to eight dihedral angles (ϕ, ψ). The PBs m and d are mainly associated to the central region of α-helix and the central region of β-strand, respectively. (de Brevern, 2005 ; de Brevern, et al., 2000).

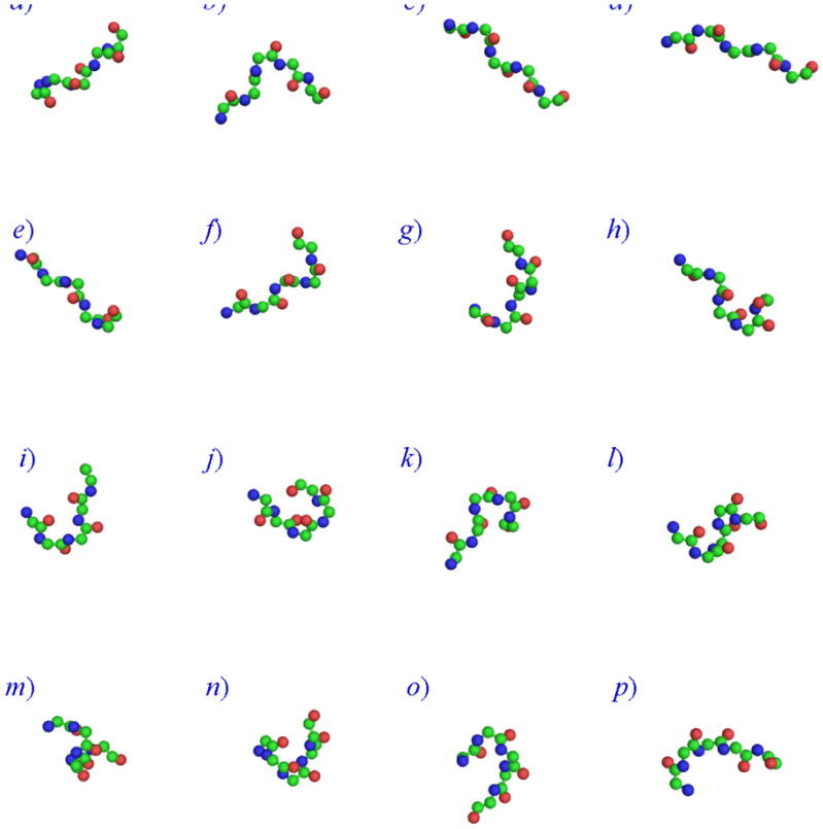

Figure 3
An example of difficult superimposition of 3D protein structures using PBs
The 3D structure of Aspergillus niger acid phosphatase (Kostrewa, et al., 1999) chain B has been superimposed on the 3D structure of Escherichia coli periplasmic glucose-1- phosphatase chain A (Lee, et al., 2003). (a) and (b) using previous approach (Tyagi, et al., 2006), and (c) and (d) with the novel approach. Using regions of high similiarity as seeds (blue, gold and pink) seen in (e), the root mean square deviation is 17 Å lower than the value previously computed.

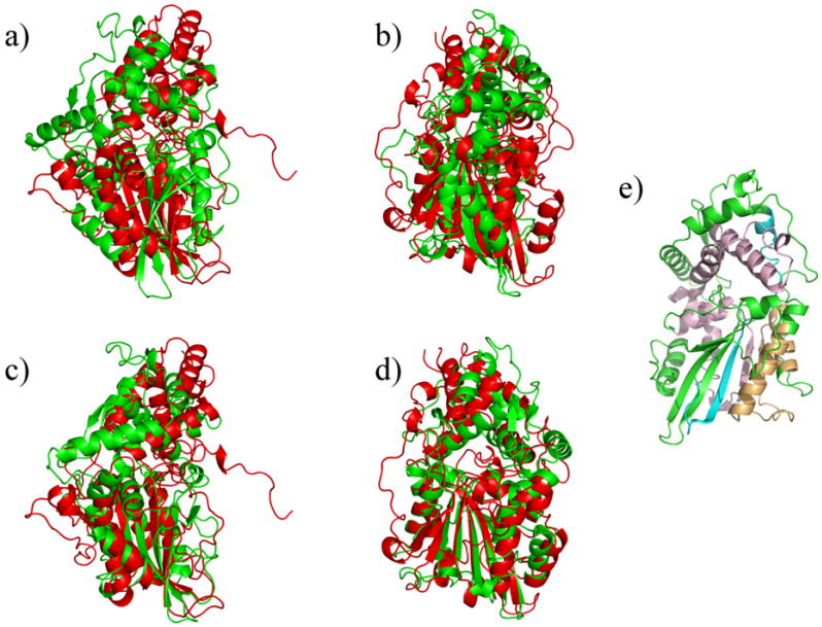

Figure 4

An example of hard case of superimposition of 3D protein structures using PBs

The 3D superimposition shown in Figure 3 is presented here as the PB alignment. Repetitive PBs and direct Ncap and Ccap are shown in colour.

Table 1

Summary of prediction methods

Given are the kind of approaches with the year and refs of publication, the prediction rate and the remarks.

approach	year	information	prediction rate (%)	Refs	web server	remarks
Bayesian prediction	2000	one sequence	34.4	de Brevern et al ., Proteins	LocPred	first method
Sequence families	2000	one sequence	40.7	de Brevern et al ., Proteins	LocPred	based on Bayesian prediction
Bayesian prediction	2002	one sequence	34.4	de Brevern et al ., Prot Sci	none	prediction of Structural Words
Hidden Markov Model	2003	one sequence	Not	Karchin et al ., Proteins	None	fold recognition
Sequence families	2005	one sequence	48.7	Etchebest et al ., Proteins	LocPred	improved Sequence Families
Pinning strategy	2007	one sequence	43.6	de Brevern et al ., J Biosc	none	prediction of Structural Words
knowledge-based prediction	2007	one sequence	62.0	Offmann et al ., Cur Bioinf	pb_prediction	pentapeptide match/SCOP class
Two-layer SVM	2008	evolutionary	61.0	Zimmermann and Hansmann, J Chem Inf Model	LOCUSTRA	first use of evolutionary information
Database-matching approach	2009	one sequence	45.3	Li et al ., Proteins	LSSRAP	use also accessibility and secondary structure
svmPRAT	2009	evolutionary	68.9	Rangwala et al ., BMC Bioinformatics	svmPRAT	protein residue annotation toolkit